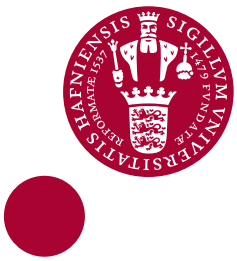


Computational redesign of thioredoxin is hypersensitive towards minor conformational changes in the backbone template

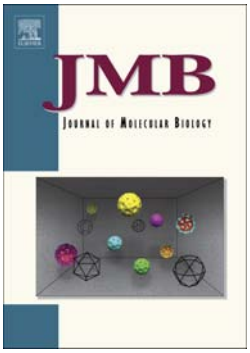
Johansson, Kristoffer Enøe; Johansen, Nicolai Tidemand; Christensen, Signe; Horowitz, Scott; Bardwell, James C. A. ; Olsen, Johan Gotthardt; Willemoës, Martin; Lindorff-Larsen, Kresten; Ferkinghoff-Borg, Jesper; Hamelryck, Thomas Wim; Winther, Jakob R.

Published in:
Journal of Molecular Biology

DOI:
10.1016/j.jmb.2016.09.013

Publication date:
2016

Document version
Peer reviewed version



Computational redesign of thioredoxin is hypersensitive towards minor conformational changes in the backbone template

Kristoffer E. Johansson, Nicolai Tidemand Johansen, Signe Christensen, Scott Horowitz, James C.A. Bardwell, Johan G. Olsen, Martin Willemoës, Kresten Lindorff-Larsen, Jesper Ferkinghoff-Borg, Thomas Hamelryck, Jakob R. Winther

PII: S0022-2836(16)30378-3
DOI: doi: 10.1016/j.jmb.2016.09.013
Reference: YJMBI 65208

To appear in: *Journal of Molecular Biology*

Received date: 5 June 2016
Revised date: 8 September 2016
Accepted date: 14 September 2016

Please cite this article as: Johansson, K.E., Johansen, N.T., Christensen, S., Horowitz, S., Bardwell, J.C.A., Olsen, J.G., Willemoës, M., Lindorff-Larsen, K., Ferkinghoff-Borg, J., Hamelryck, T. & Winther, J.R., Computational redesign of thioredoxin is hypersensitive towards minor conformational changes in the backbone template, *Journal of Molecular Biology* (2016), doi: 10.1016/j.jmb.2016.09.013

Computational redesign of thioredoxin is hypersensitive towards minor conformational changes in the backbone template

Kristoffer E. Johansson[a,†,1], Nicolai Tidemand Johansen[a,†,2], Signe Christensen[a,3], Scott Horowitz[b], James C. A. Bardwell[b], Johan G. Olsen[a], Martin Willemoës[a], Kresten Lindorff-Larsen[a], Jesper Ferkinghoff-Borg[c], Thomas Hamelryck[d] and Jakob R. Winther[a]

[a]**Linderstrøm-Lang Centre for Protein Science**, Department of Biology, University of Copenhagen, Ole Maaløes Vej 5, DK-2200 Copenhagen, Denmark

[b]**Howard Hughes Medical Institute**, Department of Molecular, Cellular and Developmental Biology, University of Michigan, 109 Zina Pitcher Place, Ann Arbor, MI 48109, USA

[c]**Biotech Research and Innovation Centre**, University of Copenhagen, Ole Maaløes Vej 5, DK-2200 Copenhagen, Denmark

[d]**Section for Computational and RNA Biology**, Department of Biology, University of Copenhagen, Ole Maaløes Vej 5, DK-2200 Copenhagen, Denmark

Correspondence to Jakob Winther: E-mail: JRWinther@bio.ku.dk. Postal address: Department of Biology, University of Copenhagen, Ole Maaløes Vej 5, DK-2200 Copenhagen, Denmark. Phone: +45 3532 1500. Fax: +45 3532 2128.

Abstract

Despite the development of powerful computational tools, the full-sequence design of proteins still remains a challenging task. To investigate the limits and capabilities of computational tools, we conducted a study of the ability of the program Rosetta to predict sequences that recreate the authentic fold of thioredoxin. Focusing on the influence of conformational details in the template structures, we based our study on 8 experimentally determined template structures and generated 120 designs from each. For experimental evaluation, we chose 6 sequences from each of the 8 templates by objective criteria. The 48 selected sequences were evaluated based on their progressive ability to: (1) produce soluble protein in *Escherichia coli*, (2) yield stable monomeric protein, and (3) the ability of the stable, soluble proteins to adopt the target fold. Of the 48 designs, we were able to synthesize 32, 20 of which resulted in soluble protein. Of these, only two were sufficiently stable to be purified. An X-ray crystal structure was solved for one of the designs, revealing a close resemblance to the target structure. We found a significant difference between the eight template structures to realize the above three criteria despite their high structural similarity. Thus, in order to improve the success rate of computational full-sequence design methods, we recommend that multiple template structures are used. Furthermore, this study shows that special care should be taken when geometry optimizing a structure prior to computational design when using a method that is based on rigid conformations.

Keywords

Computational protein design; de novo protein design; Rosetta; protein folding; protein stability

Introduction

The ability to routinely design new functional proteins and protein-based systems will significantly impact the development of novel technologies and medicinal products as well as our basic understanding of proteins. One of the major challenges in this regard is the ability to rationally design an entire amino acid sequence that will adopt a given three-dimensional structure. To handle the vast complexity of full-sequence design, computational methods are particularly interesting. Analytical or non-computational approaches have successfully been applied to the full-sequence design of α-helical structures[1—5] for example by using heptad repeats.[6,7] Also, small and less regular structures have been designed by non-computational consensus approaches and fragment assembly.[8,9] However, designing larger (>70 amino acids) globular αβ proteins with irregular contact patterns is a highly complex task and has only been achieved by employing computational methods.[10—14] In addition to this unique achievement, computational methods have been employed to full-sequence design of a variety of protein structures including early mini-proteins,[15—17] tandem repeats,[18,19] and ligand binders.[20-21] Despite much effort, however, the total number of full-sequence designed proteins for which an atomic resolution structure has been solved still remains low; to our knowledge, less than 10 larger globular αβ proteins have been reported in the literature.[10,12—14,22] Among the available computational methods used here, Rosetta is by far the best validated and thus we base this study on the Rosetta software.

This relatively low number of successful designs highlights the need for further development of computational protein design methods. To our knowledge, all computational methods capable of optimizing an entire amino acid sequence of more than 100 residues approximate side-chain degrees of freedom by a discrete, typically small number of rigid conformers referred to as rotamers[23] and the backbone is kept completely fixed during sequence optimization. We will refer to this setup as the use of rigid conformations. Today, most design protocols optimize sequence and conformation iteratively. However, in the sequence optimization step the back-

bone and rotamer conformations are always fixed. With a vast number of sequence combinations to be explored, the use of rigid conformations greatly reduces the complexity of the sequence optimization.

While being a key enabling factor in terms of computational time, the use of rigid conformations is also considered to be the main factor limiting accuracy and, in practice, it limits the application to relatively rigid proteins.[24--27] In particular, the appearance of molten globule characteristics in non-successful designs have been associated with a lack of tight packing in the hydrophobic core caused by the use of rigid conformations.[10,28] To achieve a more accurate comparative computational evaluation of structures it is necessary optimize the geometry using all degrees of freedom.[29]

When based on a single template structure, design methods based on rigid conformations are known to converge to a narrow distribution of sequences.[11] In contrast, using more templates that display minor conformational differences increases the sequence variation of the output drastically.[30,31] Together, this indicates that computationally designed sequences based on a single rigid backbone template will only result in a small subset of the sequence solution space, as defined by the applied energy function, while another template of the same fold will yield another subset of solutions even assuming the same energy function.

To investigate this, we have conducted a full computational design study in which designed sequences based on several template structures were experimentally evaluated in an unbiased fashion. The thioredoxin fold was chosen as design target because this fold is both highly conserved throughout evolution and is also realized by a large variety of sequences in nature. Thus, we expect the thioredoxin fold to have a large sequence solution space and to be highly designable in the sense that many sequences should be able to assume its fold. Furthermore, thioredoxin is a relatively rigid protein that is composed almost completely of segments with defined secondary structure (>90%) and has previously been shown to behave well in engineering contexts.[32—34] With a diversity of native sequences available, we tested templates with minor conformational changes (C_α RMSD < 2

Å), both representing natural variation resulting from different wild-type sequences, and a generated conformational variation resulting from computational geometry optimization. Starting with eight experimental template structures of the thioredoxin fold we found a significant difference between, not only the sequence outputs but interestingly also the performance in experimental evaluations from template to template. In line with previous studies, we attribute these differences in template performance to the use of rigid conformations and show that these effects are enhanced by conducting thorough geometry optimization, prior to design.

Results and Discussion

Design templates

To find a suitable set of templates to represent the most of the natural thioredoxin sequence space, we searched the PDB for structures of the thioredoxin fold that shared high structural similarity despite low similarity in amino acid sequence. To enable direct comparison of equivalent sequence positions in the resulting designs, only sets with a gap-free alignment were considered. The search resulted in eight structures, which were truncated to the common most structured 104 residues (Table 1). The structures are highly similar in backbone structure (Fig. 1) with an average C_α RMSD of 1.2 Å (0.7—1.8 Å), but diverse in amino acid sequence with an average pairwise identity of 33% (15—61 %, Fig. 2). Although it should in principle not matter, we note that all structures have been determined from protein expressed in *E. coli*, the same host that we used for expression here.

Prior to computational design of an experimental structure, the Rosetta manual recommends preparing a structure including a geometry optimization in an effort to remove disagreements between the experimental

structure and the energy function both of which potentially contain inaccuracies*. The motivation is that a disagreement between the structure and energy function, for example an atomic overlap, would favor any change that removes this overlap thus resulting in a bias away from the wild-type amino acid. By geometry optimizing the structure, an optimal energy of the wild type is achieved resulting in a more fair comparison of energies.

In this work we use the RosettaRelax application for geometry optimization. This application contains a stochastic element and therefore converges to a slightly different structure and energy in each run. Thus, five independent geometry optimizations were generated for each of the eight experimental structures resulting in a total of 48 template structures. The conformational variation between the 40 geometry optimized structures is similar to the native variation with an average pairwise C_α RMSD of 1.2 Å. However, optimizations of the same native structure are more similar (average C_α RMSD 0.5 Å) than optimizations of different structures (average C_α RMSD 1.4 Å). Otherwise, the structural differences are distributed approximately homogenously over the geometry optimizations and sequence positions.

Template assessment of design-ability

More attempts were made to assess the usefulness of the eight template structures in a design context (Table 2). Ideally, an energy function should have an experimental structure represented as a local energy minimum, in contrast, a large structural distortion upon geometry optimization may indicate a poor match between a structure and an energy function.[35] Also, the relative energy at which optimizations converge may be informative. For our set of eight experimentally determined thioredoxin structures, the average structural distortions range between 0.5 Å and 1.0 Å RMSD (Table 2). We note that the lowest resolution X-ray structure and two

* Rosetta documentation on structure preparation: https://www.rosettacommons.org/manuals/archive/rosetta3.5_user_guide/dd/da1/preparing_structures.html

NMR structures are slightly more distorted than the remaining structures. Geometry optimizations converged approximately to the same energy of -230 Rosetta energy units (REU) except for 3GNJ, 1DBY and 2L4Q. The structure 3GNJ converged to a significantly lower energy than the other structures and may thus appear to be a more promising design template while the two NMR structures converged at higher energies than all the others, making them less promising (Table 2). Since we wanted to investigate a diversity of templates, we decided to proceed to the design phase using all eight structures including even those that may appear less promising from our geometry optimizations.

Calculating the percent of side-chain conformations that can be reproduced in a repacking experiment has been suggested to provide another measure of template design-ability.[36] In contrast to the original report of this test, the application to our thioredoxin template set resulted in no significant correlation with structure resolution (Table 2). However, for geometry optimized templates, more than 99% of the side-chain conformations were reproduced in the repacking experiments which show that the minor conformational changes of the optimization were sufficient to make the same rotamer library match the structure completely.

The repacking experiment is sensitive towards the size of the applied rotamer library so we used this experiment to evaluate this in the context of our thioredoxin redesign. Rosetta allows the inclusion of additional rotamers based on the dihedral angle standard deviation given in the original description of the rotamer library. Including more samples of a rotamer mode could possibly mean the difference between reproducing the experimentally determined conformation or not. On the other hand, due to combinatorial explosion, this addition to the rotamer library is limited by computer time and memory. In the current work we were able to include two extra conformations per rotamer mode positioned at plus and minus one standard deviation for χ_1 and, for aromatic side chains, χ_2 dihedral angles.

Computational designs

For each of the 48 template backbones, we generated 20 computational designs resulting in a total of 960 designs. The nomenclature used to designate these is dNyxx; d for 'design', N for the first letter in the PDB id code (two letters may be used here in case of ambiguity), y for the geometry optimization run (1—5 and zero for non-optimized) and xx for the design run (01—20).

For each template, the 20 designs converged to a relatively narrow distribution of energies and sequences as expected (data not shown). However, sequence populations resulting from different templates were far less similar despite the close structural relationship of the templates. As a result multiple sequence alignment was able to cluster the pool of 960 designs precisely according to template origin using only the information contained in the amino acid sequences (Fig. 3). The 120 designs based on one experimental template and its geometry optimized templates (large clusters in Fig. 3) have, on average, 50 % pairwise sequence identity compared to only 30 % average identity to designs based on other experimental templates. Sequences based on geometry optimized templates of the same experimental structure are clearly separated and collected into larger clusters for each experimental template. This observation shows that, in all cases, minor conformational changes in the backbone template change the population of resulting amino acid sequences significantly and that a given template only reveals a part of the sequence solution space.

The sequence alignment further shows that the geometry-optimized templates result in slightly less diverse sequence populations (average pairwise sequence identity 68%) compared to the experimental templates (average pairwise sequence identity 61%). This trend in diversity is far more pronounced in how well the native amino acids are reproduced: comparing the individual sequence populations to its wild-type sequence shows that on average 42% of the native amino acids are reproduced for geometry-optimized templates whereas only an average of 29% are reproduced for experimental templates. This finding confirms that the intended effect of

the initial geometry optimization is substantial and results in the reproduction of more wild-type identities. However, in the final section of this paper, we present arguments that suggest this preference for the wild-type sequence may be artificially biased.

Selecting sequences for experimental characterization

In selecting sequences for experimental characterization, we rely on objective criteria. We were unable to identify a computational assessment measure, such as unsaturated and buried hydrogen bond donors and acceptors, packing statistics or a linear combination of these that was able to convincingly isolate a population of designs as more promising (data not shown). Visual inspection of randomly selected designs did not reveal any obvious problems such as hydrophobic or highly charged patches on the surface. The naturally conserved Pro 73 supports a cis peptide bond that is not always reproduced in our designs. However, a mutation to Ala, which was found in the majority of the designs that did not have a Pro, has been observed experimentally to enhance refolding properties.[37] The fact that it inactivates the enzyme[38] is not relevant for this study.

Thus, we opted for a simple objective criterion and selected, for each of eight templates, the six designs with lowest RosettaDesign energy resulting in a total of 48 sequences for experimental characterization. Interestingly, among all 960 designs the 60 best scoring all originate from 3GNJ and thus represent a minor fraction of the sequence solution space shown in Fig. 3. By selecting uniformly from all templates, our sequences for experimental characterization represent a more diverse set.

All of the 48 selected designs are based on geometry-optimized templates, and for three of the eight templates, designs are based on a single geometry-optimized template (dL4xx, dTr4xx, and dG2xx). Within each individual group, the energies of the six best scoring designs are all within less than 2 REU, which we judge to be below the general noise level.

Production in *E. coli* and solubility screen

We added a leading Met and a C-terminal His_6 tag (for purification) to each of the 48 sequences selected for experimental characterization. Plasmids containing codon-optimized sequences were custom synthesized, and the genes were expressed in *E. coli* at 37°C under the control of an IPTG-inducible T5 promoter, followed by cell lysis and centrifugation.

To evaluate expression and solubility, we performed western blots using an anti-His_5 antibody to enable detection of the expressed protein in the pellet and supernatant fractions (Fig. 4). Protein from the same amount of cells were loaded in each lane, allowing accurate determination of the relative amount of soluble and insoluble tagged protein. To quantify levels, the band intensities were compared to the intensity of a His_6-tagged control protein in three 10-fold dilutions starting from 18 pmol, which was assigned a value of 100 (Fig. S1). Expression levels span roughly 3 orders of magnitude.

Of the 48 tested designs, 16 were not produced in E. coli to a detectable level (Fig. 4, absent bars). Of the remaining 32 designs, 20 were found in the supernatant to at least some extent (Fig. 4, black bars), suggesting that they were soluble and potentially folded. In repeated experiments, some of the 16 designs, that were initially found to be not produced, were found to express sporadically, but never with a significant yield of soluble protein.

A key issue in full sequence design is to obtain significant amounts of soluble protein for further characterization and optimization. Without soluble protein, there is little basis for further efforts. To this end, we designated a design as promising if the soluble fraction was greater than 1 (as defined in Fig. 4). This resulted in a total of 9 promising designs out of the 48 tested. Given this total statistics, the binomial probability of obtaining 6 promising designs out of 6 tested (as for 2I4A) in 8 attempts is $8 \times \left({}^{9}/_{48}\right)^{6} \sim 0.03\%$, making it unlikely that the template dependency shown in Fig. 4 is a coincidence. The solubility screen provides

strong evidence that the outcome of a computational design is sensitive to minor conformational changes in the backbone template. We note that the success rate in the solubility screen is vastly better than the null hypothesis since random sequences of 100 amino acids, for all practical purposes, are expected to be insoluble.[39]

The solubility screen suggests that three templates (3GNJ, 1FB0, and 2I4A) are more successful than the others. Thus, we evaluated the computational template assessment (Table 2) based on this solubility data (Fig. 4). The most promising template by energy was 3GNJ, the most promising by χ_1 reproduction was 1FB0, and the most promising by X-ray resolution was 2I4A. However, we found no measure that could consistently rank all templates in accordance with the solubility screen results. The energy and distortion upon geometry optimization suggest that the two NMR structures (2L4Q and 1DBY) and to a lesser extent the low-resolution structure 3HZ4, do not perform well with Rosetta.

Among the six designs on each template, we found no correlations between solubility and design energies or other computational post-evaluations, such as geometry-optimized Rosetta energies, that were able to predict the more successful of the six designs (data not shown).

Purification and monomer stability

We attempted to purify the most promising soluble designs using immobilized nickel affinity chromatography. Most of these proteins either did not elute from the column, presumably because of on-column aggregation or were not stable or soluble enough to stay in solution after this initial purification step. To ensure that proteins that passed this step were not significantly multimeric, they were subjected to size exclusion chromatography. Only two designs, dF106 and dF414, were stable enough to be purified using size exclusion chromatograhy. Of the two, dF106 showed better solubility but only at low pH. Interestingly, we predicted the pI of dF106 to be 4.8, which suggests that charge neutralization may be required to keep dF106 soluble. The low solubility of

dF414 (~30 μM) was maintained over a broader pH range (from 4 to 10). Consequently, dF414 was analyzed in near neutral pH buffers, whereas dF106 was analyzed in pH 4.8 buffers.

In the size exclusion chromatography, both dF106 and dF414 could be recovered from peaks eluting close to the expected retention volume, suggesting that these two proteins are monomeric in solution and reasonably compact (Fig. 5a). The slightly reduced retention time of dF106 could be indicative of it having a slightly extended structure. If kept at room temperature for several days, however, dF106 showed a significant amount of multimer formation (Fig. 5a, black curve). The peak eluting at 11.8 mL translates to a molecular mass corresponding to a trimer of dF106. A small peak is also visible around 13 mL for the freshly prepared dF106 sample (Fig. 5a, red curve). For practical reasons, the gel filtration experiment was run at room temperature, which could have caused minute dimer formation to occur. In the purification of dF106, all protein was collected from the monomer peak. For both dF106 and dF414 the recovered protein was >95% pure as determined by SDS-PAGE (data not shown).

Structural analysis

We biophysically characterized the two stable and monomeric designs, dF106 and dF414. Far ultraviolet (UV) circular dichroism (CD) spectra indicate an αβ structure with the correct amount of each secondary structure type (Fig. 5b), and the near UV spectra are indicative of well-defined tertiary structures (Fig. S2). The structures of the two designs appear to be very resistant to thermal denaturation (Fig. S3) as is commonly observed for computationally designed proteins[1,2,4,10–12,14]; dF106 started unfolding at ~70°C, whereas the CD spectra of dF414 barely responded to the heat treatment. Both designs were responsive to chemical denaturation with GuHCl, which caused significant changes in Trp fluorescence intensity (Fig. 5c) and shifts in the maximum-intensity wavelength (λ_{max}) (Fig. 5d). Whereas dF106 only has one buried Trp, dF414 has three Trp residues and

consequently a much broader spectrum; thus, the shift in λ_{max} for dF414 was much less pronounced (data not shown).

Chemical denaturation showed three-state unfolding for both designs (normalized data is shown in Fig. 5c and raw data in Fig. S4), which is in contrast to the simple two-state unfolding that is common to many small naturally occurring proteins, including the native 1FB0 template.[40] We note that complex folding mechanisms are commonly observed for designed proteins.[41,42] The unfolding of dF106 was fully reversible, but the stability of the intermediate state was dependent on the protein concentration (Fig. 5c, red and black). The intermediate state was significantly stabilized in the 5.3 μM samples compared to the 0.9 μM samples, which suggests that the intermediate may tend to dimerize or oligomerize. To test whether the data indeed indicate an oligomerization, normalized fluorescence intensities were measured in 4.5 M GuHCl samples as a function of protein concentration. These data were fitted to a simple binding curve, which suggests that the intermediate state is a dimer or oligomer (Fig. S5). The unfolding of dF414 revealed a highly stable intermediate state at the low protein concentration tested. Due to the low solubility of dF414, higher concentrations were not tested. The unfolding of dF414 was reversible in the second transition but not in the first (data not shown). This result is consistent with the observation that dF414 could not be refolded after purification from inclusion bodies (data not shown). In conclusion, the data do not suggest a simple three-state model for either of the two proteins, which consequently mean that the fitted thermodynamic parameters are not easily interpreted.

The general difference in thermal and chemical unfolding characteristics between native and designed proteins observed here and elsewhere[1,2,4,10–12,14] suggests that cooperative protein folding is an optimized property in nature that is not easily recreated in computational protein design.

In previous computational protein design studies, a common problem has been that the designed proteins resembled molten globules lacking uniquely defined tertiary structure.[43] We therefore investigated the structure

of dF106 by NMR spectroscopy. The ^{1}H-NMR spectrum of dF106 (Fig. S6) has downfield-shifted amide protons (above 8.5 ppm), well-dispersed peaks in the H_α region (around 5 ppm), and upfield-shifted methyl peaks (below 0 ppm); all of these observations indicate that dF106 has a fold with well-defined tertiary structure. Because of its low solubility, NMR spectroscopy was impractical for the structural analysis of dF414.

For dF106, we were able to obtain crystals that diffracted to 2.4 Å using purified protein that had been kept on ice for less than 2 weeks. The X-ray structure of dF106 was solved to an R_{work} = 0.20, R_{free} = 0.28 (full statistics shown in Table S1) by molecular replacement using the structure of its template 1FB0 and another thioredoxin structure, 2PUK. To check if the final structure was affected by model bias, we calculated a composite simulated annealing omit map over all copies of the final structure. The excellent fit of the structure in the composite omit map demonstrates that the final structure is essentially free of model bias (Fig. 6a).

The X-ray structure of dF106 is in excellent agreement with the computational design (Fig. 6b). The backbones of seven of the eight molecules in the asymmetric unit are fully resolved, except for the His_6 tag, and align to the design model with a C_α RMSD of 1.8–2.0 Å. The structural difference between the X-ray structure and the design model primarily originates from a displacement of the N-terminus and N-terminal helix (positions 1–16) (Fig. 6b, right). This helix contains a surface exposed hydrophobic patch consisting of four Leu side chains (positions 10, 11, 14, and 15). This patch constitutes the main packing contact of the crystal complex implying that crystal contacts may contribute to the distortion seen from the design structure. Another noticeable difference is observed at the beginning of the long helix 2 (Fig. 6b, top left). Wild-type thioredoxin features a conserved kink in the long helix generated by Pro 37, which is thought to be functionally important.[44] In the design, where function was not a design criteria, Pro 37 is not present and, as a result, the encompassing helix lacks this kink. Excluding the 16 N-terminal C_α atoms and those distorted by Pro 37 (res 26–33) results in a RMSD of 0.7–0.9 Å

to the template. Chain D has the lowest energy after geometry optimization of -260 REU which is notably larger than the -275 REU of the design.

The X-ray structure shows that most of the designed contacts are realized. In non-surface positions, 85% of χ_1 dihedral angles are within 40° (Fig. 6c shows examples). An interesting exception is Val 2 which forms the contact to the hydrophobic core intended for Val 4 due to the displacement of the N-terminal residues. Side-chain χ_1 angles, that are not realized, are situated near the absent Pro 37 helix kink (Asp 29, Val 35), in the N-terminal helix (Val 4 and Leu 12), or have little impact on the position of side-chain atoms (Leu 96 and Arg 97). The lack of designed contacts in the displaced N-terminal helix could be the origin of the slightly reduced retention time and slow oligomerization of dF106 observed experimentally (Fig. 5a).

Sequence post-analysis

We examined the success of dF106 and dF414 at the sequence level by comparing these two designs and the wild-type sequence of the template structure 1FB0 with the other 30 sequences that had been expressed and found to be either soluble or insoluble (Fig. 7).

The backbone conformation of a position alone may in some cases have a high influence on the design outcome of that position. For example, position 89 in all templates populates a region of the Ramachandran plot that only allows Gly (ϕ = 95 ± 15°, ψ = 180 ± 30°) and thus Gly 89 is invariably reproduced in our designs. Val and Ile, known to have a high propensity for β-sheet structure, are reproduced in the central sheet at positions 20, 22, 52, 75, and 88 in our designs. Compared to Gly 89, we expect non-bonded as well as bonded energy terms to contribute here. Other conserved core positions include Val 13, Phe 24, Leu 55, Thr 74, Phe 77, and Phe 78, resulting in a highly conserved hydrophobic core among the designs selected for experimental characterization. However, since these positions are conserved in both the successful and non-successful designs (Fig. 7), no discriminative power can be assigned to this observation.

A buried hydrogen bond is formed between the side chains of Trp 9 and Tyr 67 in the native structure of 1FB0 and this is reproduced in the successful designs dF106 and dF414, but rarely in the non-successful designs (Fig. 7). In the wild-type sequence of the other templates, one or both of these residues are Phe, which is also the case for the other tested designs (Figs. 2 and 7). Other positions also appear to correlate with solubility and/or stability, for example Met 21, Gly 48, Glu 82, and Lys 84, but for less obvious reasons.

The wild-type Asp 23 is highly conserved in native thioredoxins despite a thermodynamically unfavorable position in the hydrophobic interior of the protein. This has been shown experimentally by a significant pK_a shift of the carboxyl acid to 7.5.[45] Both dF106 and dF414, in contrast to most other tested designs, have the isostructural Leu at position 23, which may contribute significantly to their success. Surprisingly, some designs reproduce the buried Asp, which we will discuss in the last section of the paper.

Reproduction of natively conserved positions

Since a significant fraction of the wild-type identities were reproduced, we were able to identify the template origin of each design from its sequence alone (Fig. 3). Thus, it is interesting to explore the correlation between naturally conserved positions and their template dependent reproduction in the computational designs (Fig. 8 for 1FB0 and Figs. S7–S14 for all templates). Several positions in wild-type thioredoxins have previously been assigned as conserved for either a structural purpose (23 red positions in Fig. 8) or for a functional purpose (14 green positions in Fig. 8).[44] In all 960 designs, reproduction of the structural conserved positions is only slightly greater (41%) than reproduction of the functionally conserved positions (36%). The latter slight decrease in reproduction of functionally conserved positions is almost entirely accounted by the observation that the redox active CGPC loop was rarely reproduced in the designs. On average, 39% of the wild-type amino acids were reproduced in the designs. Thus, naturally conserved amino acids are not reproduced more often than other residues in the computational designs. This is somewhat surprising since the energy function should recognize

residues that are important for the fold but not those related to the function. In the following section we will investigate the mechanism of residual reproduction in the computational design method.

The eight experimental template structures do not always contain the naturally conserved amino acids. In these cases, the reproduction in design follows the template rather than the natural consensus. For example, the wild-type sequence of 1FB0 does not have the naturally conserved Phe 9, Leu 21, Met 34, Asn 60, and Val 88 residues, and in these cases, the computational design tends to reproduce the template identity rather than the naturally conserved identity (Fig. 8). The naturally conserved but thermodynamically unfavorable Asp 23 is reproduced in 14% of designs, but interestingly not in designs based on the successful template 1FB0. This reproduction of Asp 23 is highly template dependent and is explored further in the following section.

Computational investigation of template dependency

To investigate the apparent high sensitivity towards minor conformational changes in the template, we conducted five additional in silico experiments, based on the observation that two templates, dTr4xx and dT2xx, tend to reproduce the buried Asp 23 (Figs. S12 and S9 respectively). This buried Asp is reproduced despite it having a thermodynamically unfavorable position evident from the significant pK_a shift of the carboxylic acid to 7.5.[45] All experimentally evaluated designs that were based on the two templates dTr4xx and dT2xx, were found to be insoluble or unstable. In contrast, the two successful designs, dF106 and dF414, both have the isostructural Leu at position 23 (Fig. 7). We therefore conducted the following experiments to explore Leu as a better choice at postion 23. In these experiments, we generated 100 new designs to enhance the statistics from the 20 designs generated previously. We monitored the average positional energy of Asp and Leu under the approximation of rigid conformations (termed packer energy) and for a post-design geometry-optimized structure (termed relax energy). The latter is known to correlate better with stability.[29]

In experiment A, we confirmed the reproduction of Asp 23 by these two templates. In experiment B, position 23 was fixed to Leu to obtain energies for comparison. These data show that the original design settings mainly resulted in Asp at position 23 due to the use of rigid conformations even though the relax energies indicate that the energy function recognized Leu as the better choice in the optimized geometry (Table 3, rows A and B). However, the energy of position 23 varies 1–2 REU between individual designs with the same identities, which makes the average energy difference insignificant and following, we expect a population close to 1:1 of Asp and Leu with knowledge of the relaxed energies. In experiment C, we used the highest number of rotamers available in Rosetta (12 additional conformations per rotamer) for position 23. This resulted in a minor increase in the preference for Leu in 2TRX (from 2% to 8%), but no increase with 1T00. For both templates, Asp was still preferred. In experiment D, we used the more recent Talaris2013 energy function[46], most notably featuring a Coulomb type of electrostatic energy term. This destabilized the Asp significantly, resulting in only 15% occurrence for 1T00 and 0% for 2TRX. However, the Asp was to a large extent replaced by Asn or Ala and not, as expected, by a larger hydrophobic side chain like Leu. In order to compare energies, we optimized these designs with the score12 energy function used in all other aspects of this work. Finally, in experiment E, we made a new template by mutating position 23 to Leu followed by geometry optimization and design using the original settings of experiment A. Interestingly, this resulted almost exclusively in Leu implying that the sequence used during geometry optimization strongly biases the outcome as it is also seen in the initial repacking assessment of the templates (Table 2).

These experiments show that subtle conformational changes to the backbone template may be far more significant for the design outcome than the size of the rotamer library or choice of energy function in Rosetta. As a result of the initial geometry optimization used in this study, the unfavorable buried Asp 23 was reproduced with two specific templates. Probably due to the inclusion of a Coulomb term, the use of the Talaris2013 energy function largely avoided reproduction of the buried carboxylic acid, but only when the template was opti-

mized with Leu at position 23 did the design result unambiguously in a hydrophobic side chain larger than Ala. Inspection of individual energy components showed that non-bonded energy components dominate the results in Table 3, which consistently indicates that the approximation of rigid conformations was the primary source of the reproduction of the unfavorable Asp 23.

The approximation of rigid conformations is commonly considered to be one of the main limitations to accuracy and iterative optimization of sequence and backbone conformation has been reported as critical to the success of computational design.[10] To further analyze these suppositions, we applied the RosettaRelax and RosettaDesign applications iteratively ten times for the two templates dTr4xx and dT2xx. After ten cycles, the energy and sequence changed very little and the iterative design protocol was deemed to have converged. Again, both experiments resulted in reproduction of Asp 23. For one run, the first iteration suggested Met at position 23, but even then, in subsequent rounds this reverted back to Asp.

The case of Asp 23 highlights the risk of using thorough geometry optimization in the preparation of a design template. A conservative geometry optimization protocol seeks to reduce atomic displacements[47]; however, this assumes that reproduction of the native amino acids is always desirable. The large pK_a shift of Asp 23 in the native structure of thioredoxin and the observation that both of our successful designs have a Leu at position 23 strongly indicates that reproduction of native amino acids is not necessarily a good target for a designed sequence. The results in Table 3 further indicate that the reproduction of Asp 23 is an artifact caused by the approximation of rigid conformations and not representing any physical aspects of the model. Specifically, the relaxed energies suggest that the energy function recognizes Leu as the better choice but these energies are not available in the sequence optimization step which is based on rigid conformations. Only by adapting the two templates to Leu in experiment E enables the sequence optimization method to recognize the more stable

Leu. The mechanism of artificial reproduction is thus an over-consistent match between the optimized backbone template and the rotamer library rather than inaccuracies in the energy function.

Can we, based on our observations, retrospectively predict which template(s) would be most promising? The answer is somewhat disappointing. None of the independent computational metrics applied clearly suggested 1FB0 as a superior template. The relative energy of designs based on different templates is not informative since the 60 overall best scoring designs are all based on the template 3GNJ, which did not yield any folded protein. In addition, the very practical solubility and expression screen (Fig. 4) also had little predictive power in its own right and 1FB0 was not the best scoring template by these criteria. We would like to note that the consistency with which *E. coli* distinguishes different templates, in spite of the fact that designs within the same template are only in the order of 50% identical, is surprising and obscure.

At this point we can make three recommendations: A) Multiple templates should be tested. B) The preference for wild-type identities, discussed above, suggests that templates should be selected as diverse as possible in wild-type sequence. C) In case of sparse resources, a template may be discarded if it scores significantly worse in more assessments, as is the case for 1DBY (Table 2).

Concluding remarks

We have shown that the success rate in redesigning the thioredoxin fold using Rosetta is highly dependent on minor conformational changes in the backbone template (C_α RMSD < 2 Å). We tested eight experimental structures with high structural similarity and low sequence similarity and found that one template resulted in two stable monomeric proteins. We were able to solve the structure of one of these designs using X-ray crystallography. Five of the eight template structures only resulted in poorly soluble protein in low yield. Three of these poor templates could to some extent be identified by objective computational criteria, but we were not able to identify the successful template 1FB0 without experimental assessment. Thus, we conclude that in computa-

tional design experiments it is advantageous to use multiple templates and include all in the experimental evaluation.

We found that the reproduction of template wild-type amino acids is artificially enhanced when a template structure is geometry-optimized using RosettaRelax prior to design and link this to the use of rigid conformations in the sequence optimization. As in previous studies, our findings indicate that the use of rigid conformations is a major limitation to the accuracy of computational protein design.

Materials and Methods

Computational

The PDB was searched for homologs of the *E. coli* thioredoxin sequence, resulting in a list of 283 PDB entities. The identification of a suitable set of templates from this list was automated in a computer program based on the BLAST stand-alone tools version 2.2.23[48] and the BioPython software library version 1.58.[49] This was the largest set of structures we were able to obtain that met our conditions (high structural similarity, low sequence similarity, and gap-free alignment). For X-ray structures with more than one chain in the asymmetric unit we always used the first chain and for NMR structures we always used the first model. The quality of the X-ray structures was assessed by comparing C_α position deviations between the structures deposited in the PDB and those deposited in the PDB_REDO database.[50] The structures deviated 0.02–0.13 Å in C_α RMSD. The structure 2TRX did not have deposited structure factors and could not be assessed for quality. The two NMR structures, 2L4Q and 1DBY, were assessed by ResProx[51] to have an equivalent resolution of 1.1 Å and 1.7 Å, respectively.

We used the RosettaDesign (fixbb) and RosettaRelax (fast_relax) applications from Rosetta version 3.1 for sequence and structure optimization. Unless stated otherwise, all work is based on the score12 energy function

and Dunbrack 2002 rotamer library.[52] For protein design, we used the protocol described by Dantas et al.[11] except that the initial reduction of the search space was omitted so that the entire sequence space was searched with extra rotamers. We choose this protocol because it is the best validated for redesign of a native fold. Designs based on 3GNJ and 3HZ4 retained the active disulfide, resulting in an additional stability of ~8 REU, which was subtracted from all reported energies. However, this did not change rankings or conclusions. In the template repacking test, the native sequence was fixed and side-chain conformations were considered reproduced if the first dihedral angle (χ_1) of flexible and buried side-chains were within 40° of the experimentally determined conformation. For all templates, the following non-surface positions were considered in the repacking test: 4, 9, 12, 13, 16, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 32, 35, 38, 39, 42, 43, 46, 50, 51, 52, 53, 54, 55, 56, 57, 60, 63, 64, 67, 69, 72, 73, 74, 75, 76, 77, 78, 87, 91, 96, 97, 99, 100 and 103.

The designs reported in Table 3 that are based on the Talaris2013 energy function were generated using Rosetta 3.5 (2015.38.58158) and the Dunbrack 2010 rotamer library.[53] Geometry optimization of these designs was performed using the program and settings used for all other reported results to ensure comparable energies.

The phylogenetic tree (Fig. 3) was generated using ClustalX version 2.1[54] with the bootstrapped neighbor-joining algorithm, and the figure was made using the unroot application of NJplot.[55] Protein structures were visualized (Figs. 1 and 6) using the open-source version of PyMol[56], and graphs (Figs. 4 and 8) were made in R.[57] Sequences were visualized (Figs. 2 and 7) using WebLogo version 2.8.2.[58]

The isoelectric point (pI) was predicted for dF106 by geometry optimization using RosettaRelax followed by an electrostatics evaluation using PropKa.[59] In contrast to many other pI prediction methods, this approach accounts for pK_a shifts due to tertiary interactions in the folded protein.

Gene synthesis and protein expression

Codon-optimized genes were custom synthesized and inserted into expression vectors (pD441-CH) carrying an IPTG-inducible T5 promoter (DNA2.0). Expression was carried out in E. coli MC1061 at 37°C. Tests of protein production in E. coli BL21(DE3) at 15°C were also conducted, but yields were generally poorer than in MC1061 at 37°C and none of the designs stood out as particularly favored by this strain and temperature. Cultures containing 5 mL or 50 mL of a phosphate-buffered salt medium with the addition of tryptone and yeast extract[60] supplemented with kanamycin (30 µg/mL) were grown to mid-log phase and induced with 1 mM IPTG for 3.5 h at 37°C or overnight at 15°C. For large-scale expression, one liter cultures were grown to late-log phase and induced with 1 mM IPTG.

Western blot

After expression, cell cultures were harvested by centrifugation, resuspended in 1/25 of the culture volume (V_{start}) in lysis buffer (25 mM TrisHCl pH 8, 300 mM NaCl, 1 mM EDTA, 1 mM PMSF), and lysed by sonication while kept on ice. The lysates were centrifuged at 14100*g* for 20 min at 4°C. The top of the supernatant was carefully pipetted to a new tube to avoid pellet debris, and any residual supernatant remaining was discarded. The pellet was resuspended in V_{start} of lysis buffer. The supernatant and pellet samples were mixed 1:1 with sample buffer containing OPRTase-His_6[61], which we used as an internal standard, in three 10-fold dilutions. Equal volume samples were subjected to SDS-PAGE on 15% gels and subsequently blotted to nitrocellulose membranes. The blots were developed with a primary mouse anti-His_5 antibody (Qiagen) and a secondary rabbit anti-mouse antibody (Dako) linked to an alkaline phosphatase and developed with NBT/BCIP (Sigma).

Protein purification

After harvest of large-scale cultures, the cells were kept on ice and lysed using a French press (American Instrument). The supernatant recovered from centrifugation of the lysate was subjected to immobilized metal

affinity chromatography on NiNTA agarose (Qiagen) in Tris-containing buffers according to the recommendations of the supplier. Size-exclusion chromatography was performed using a Superdex-75 10/300 column (GE Healthcare) fitted to an Agilent 1100 HPLC system. Protein-containing fractions from the NiNTA column were pooled and 1 mL aliquots applied to the column with a flow rate of 0.5 mL/min. The void volume was ~8 mL. Buffers for size-exclusion were 50 mM NaOAc pH 4.8 (dF106) or 50 mM Tris-HCl pH 7.6 (dF415), both with 150 mM NaCl. The same procedure was used for preparative and analytical SEC.

CD spectroscopy

CD data were collected on a Jasco800 spectrometer. Far-UV CD wavelength scans (260–195 nm) at 25°C were collected in a 1 mm path-length cuvette, and near-UV CD wavelength scans (320-250 nm) at 25°C were collected in a 0.5 mm path-length cuvette, with the temperature controlled by a Peltier device. For the far-UV CD, samples contained ~0.1 mg/mL protein, and for the near-UV CD, the dF106 and dF414 samples contained ~0.1 mg/mL and ~0.37 mg/mL, respectively. The protein concentration was determined using a Specord S10 spectrophotometer (Zeiss). Protein spectra were buffer-subtracted and the CD signal was converted to mean residue ellipticity or molar ellipticity. In the temperature scans, the CD signal was recorded at 220 nm with a scan rate of 30 (dF414) or 120°C/h (dF106). Buffers were 25 mM sodium phosphate pH 7.2 and 4.8, respectively.

Fluorescence

Chemically-induced unfolding was followed by tryptophan fluorescence on a LS55 spectrofluorometer (Perkin Elmer). Samples with varying concentrations of guanidine hydrochloride (GuHCl) in 50 mM NaOAc pH 4.8 (dF106) and 50 mM phosphate pH 7 (dF414) were incubated for 24 h at 25°C. In the fluorimeter, the samples were excited at 280 nm, and the signal at 340 nm (dF106) or 348 nm (dF414) was integrated for 10 seconds. The excitation slit was kept at 15 nm and the emission slit was adjusted on the 0.5 M GuHCl sample and fixed to obtain the highest possible signal in the experiment. The obtained data sets were globally fit to a simple

three-state model.[62] The baselines were extracted and the data were converted to fraction folded with the transformation

$$f_{\text{folded}} = \frac{Y_{\text{obs}} - Y_{\text{min}}}{Y_{\text{max}} - Y_{\text{min}}}$$

where Y_{max} and Y_{min} are the baseline values of the native and unfolded state (Fig. S4).

To investigate dimer/multimer formation of dF106, the protein concentration was varied in samples containing 4.5 M and 6.0 M GuHCl, and the emission slit was adjusted to obtain the best signal. The signals in the 4.5 M samples were normalized to the 6.0 M samples and fitted to a simple binding model

$$y = \alpha + \frac{B_{\text{max}}[P]_0}{K_{\text{d}} + [P]_0}$$

where B_{max} is the maximum binding, $[P]_0$ is the total concentration of protein, and K_d is the association constant. All fitting was carried out in MatLab (MathWorks).

NMR spectroscopy

The sample for NMR contained 2 mg/mL dF106 in 100 mM sodium sulfate, 10% D_2O and DSS at a total volume of 600 µL. The pH was adjusted to pH 4.8 with acetic acid. A 1D ^{1}H-NMR spectrum was recorded on a Varian INOVA 750 Mhz (^{1}H) NMR spectrometer with a 5 mm room temperature probe at 5 and 25°C. The spectrum (Fig. S6) represents 1000 transients of 8K data points. The chemical shifts were referenced to internal DSS at 0.00 ppm. Data were processed in NMRPipe.

Crystallographic analysis

Single crystals of dF106 were obtained from hanging drops consisting of 2 µL protein solution (3 mg/mL protein in 25 mM sodium acetate pH 4.8) and 2 µL crystallization solution (10% (w/v) PEG 4000, 100 mM imidazole pH

7.0, 100–150 mM lithium citrate, 5 mM $CuCl_2$, 1 mM $(NH_4)_2SO_4$). Crystals were grown at 20°C. To obtain isolated, single crystals, the drop was streak-seeded with crushed crystals grown from identical conditions in an earlier trial. X-ray diffraction data was collected at the European Synchrotron and Radiation Facility (ESRF) beamline ID30A-3 using a micro-focus beamline with 0.1° slicing. In total, 3600 frames were recorded. The images were analyzed using iMosflm[63] and XDS.[64] The space group was determined to be $P2_12_12_1$ with unit cell parameters a = 59 Å, b = 69 Å and c = 230 Å, which suggests an asymmetric unit containing eight molecules.

To solve the crystal structure, we used the MRage automated molecular replacement pipeline in Phenix.[65] The input search models into MRage were 1FB0 and 2PUK. After molecular replacement, the top solution used for initial model building was performed with Phenix Phase and Build, followed by iterative model building and refinement using Coot[66] and Phenix Refine. To remove model bias, the first round of automated refinement after initial building included a simulated annealing refinement step. Non-crystallographic symmetries were not used in the refinement. Seven of the eight monomers in the asymmetric unit are essentially identical, with C_α RMSD's of 0.2–0.8 Å. By comparison of chain conformations and B-factors, the loop between position 25 and 33 is shown to be flexible. Disregarding this loop results in an RMSD of 0.2–0.4 Å. The only monomer to display any significant differences was partially unfolded at the C terminus, perhaps due to the uncleaved His_6 tag used for purification. With a lowest average B-factor of 80.1 $Å^2$, chain D was chosen for visualization and comparison. Crystallographic data and refinement statistics are given in Table S1.

Accession numbers

The crystal structure of dF106 is deposited in the PDB under id 5J7D.

Acknowledgements

This work was supported by the Danish Council for Independent Research grant FTP274-08-0124 and FTP0602-01373B to JRW and MW, respectively. K.L.L. was supported by the Novo Nordisk Foundation. J.C.A.B was supported by the National Institutes of General Medicine grant, R01-GM102829, J.C.A.B. is a Howard Hughes Medical Institute Investigator. We would like to thank Poul Nissen, Centre for Membrane Pumps in Cells and Disease, Aarhus University, Denmark for generously letting us use his in-house diffractometer equipment. Marianne Mortensen is thanked for excellent technical assistance and Birthe B. Kragelund for acquisition of NMR spectra.

† K. E. Johansson and N. Tidemand Johansen contributed equally to this work.

[1] Present address: K. E. Johansson, Department of Pharmacy, University of Copenhagen, Universitetsparken 2, DK-2100 Copenhagen, Denmark.

[2] Present address: N. Tidemand Johansen, Niels Bohr Institute, University of Copenhagen, Blegdamsvej 17, DK-2100 Copenhagen, Denmark.

[3] Present address: S. Christensen, Institute for Molecular Bioscience, The University of Queensland, St Lucia QLD 4072, Australia.

Abbreviations used

RMSD, root mean square deviation; REU, Rosetta energy units; PDB, protein data bank; NMR, nuclear magnetic resonance; SDS-PAGE, sodium dodecyl sulfate polyacrylamide gel electrophoresis; UV, ultraviolet; CD, circular dichroism; GuHCl, Guanidine hydrochloride; HPLC, high performance liquid chromatography; IPTG, isopropyl β-

D-1-thiogalactopyranoside; PMSF, phenylmethane sulfonyl fluoride; DSS, 4,4-dimethyl-4-silapentane-1-sulfonic acid.

References

1. Fezoui, Y., Weaver, D. L. & Osterhout, J. J. (1994). De novo design and structural characterization of an alpha-helical hairpin peptide: a model system for the study of protein folding intermediates. Proc. Natl. Acad. Sci. USA 91, 3675—3679.
2. Schafmeister, C. E., LaPorte, S. L., Miercke, L. J. & Stroud, R. M. (1997). A designed four helix bundle protein with native-like structure. Nat. Struct. Biol. 4, 1039—1046.
3. Johansson, J. S., Gibney, B. R., Skalicky, J. J., Wand, A. J. & Dutton, P. L. (1998). A native-like three-α-helix bundle protein from structure-based redesign: A novel maquette scaffold. J. Am. Chem. Soc. 120, 3881–3886.
4. Bryson, J. W., Desjarlais, J. R., Handel, T. M. & Degrado, W. F. (1998). From coiled coils to small globular proteins: Design of a native-like three-helix bundle. Prot. Sci. 7, 1404–1414.
5. Thomson, A. R., Wood, C. R., Burton, A. J., Bartlett, G. J., Sessions, R. B., Brady, R. L. & Woolfson, D. N. (2014). Computational design of water-soluble α-helical barrels. Science 346, 485—488.
6. Offer, G. & Sessions, R. (1995). Computer modelling of the α-helical coiled coil: Packing of side-chains in the inner core. J. Mol. Biol. 249, 967—987.
7. Grigoryan, G. & DeGrado, W. F. (2011). Probing designability via a generalized model of helical bundle geometry. J. Mol. Biol. 405, 1079—1100.
8. Sander, C., Vriend, G., Bazan, F., Horovitz, A., Nakamura, H., Ribas, L., Finkelstein, A. V., Lockhart, A., Merkl, R., Perry, L. J., Emery, S. C., Gaboriaud, C., Marks, C., Moult, J., Verlinde, C., Eberhard, M., Elofsson, A.,

Hubbard, T. J. P., Regan, L., Banks, J., Jappelli, R., Lesk, A. M. & Tramontano, A. (1992). Protein design on computers. Five new proteins: Shpilka, grendel, fingerclasp, leather, and aida. Proteins 12, 105–110.

9. Mayo, K. H., Ilyina, E. & Park, H. (1996). A recipe for designing water-soluble, β-sheet-forming peptides. Protein Sci. 5, 1301–1315.
10. Kuhlman, B., Dantas, G., Ireton, G. C., Varani, G., Stoddard, B. L., & Baker, D. (2003). Design of a novel globular protein fold with atomic-level accuracy. Science 302, 1364–1368.
11. Dantas, G., Kuhlman, B., Callender, D., Wong, M., & Baker, D. (2003). A large scale test of computational protein design: Folding and stability of nine completely redesigned globular proteins. J. Mol. Biol. 332, 449–460.
12. Koga, N., Tatsumi-Koga, R., Liu, G., Xiao, R., Acton, T. B., Montelione, G. T., Montelione, G. T., & Baker, D. (2012). Principles for designing ideal protein structures. Nature 491, 222–227.
13. Xiong, P., Wang, M., Zhou, X., Zhang, T., Zhang, J., Chen, Q., & Liu, H. (2014). Protein design with a comprehensive statistical energy function and boosted by experimental selection for foldability. Nat. Commun. 5, 5330.
14. Huang, P.-S., Feldmeier, K., Parmeggiani, F., Velasco, D. A. F., Höcker, B. & Baker, D. (2016). De novo design of a four-fold symmetric TIM-barrel protein with atomic-level accuracy. Nat. Chem. Biol. 12, 29–34.
15. Dahiyat, B. I. & Mayo S. L. (1997). De novo protein design: Fully automated sequence selection. Science 278, 82—87.
16. Harbury, P. B., Plecs, J. J., Tidor, B., Alber, T. & Kim, P. S. (1998). High-resolution protein design with backbone freedom. Science 282, 1462—1467.
17. Kraemer-Pecore, C. M., Lecomte, J. T. J. & Desjarlais, J. R. (2003). A de novo redesign of the WW domain. Protein Sci. 12, 2194–2205.

18. Brunette, T. J., Parmeggiani, F., Huang, P. S., Bhabha, G., Ekiert, D. C., Tsutakawa, S. E., Hura, G. L., Tainer, J. A. & Baker, D. (2015). Exploring the repeat protein universe through computational protein design. Nature 528, 580–584.

19. Doyle, L., Hallinan, J., Bolduc, J., Parmeggiani, F., Baker, D., Stoddard, B. L. & Bradley, P. (2015). Rational design of α-helical tandem repeat proteins with closed architectures. Nature 528, 585–588.

20. Cochran, F. V., Wu, S. P., Wang, W., Nanda, V., Saven, J. G., Therien, M. J. & DeGrado, W. F. (2005). Computational de novo design and characterization of a four-helix bundle protein that selectively binds a nonbiological cofactor. J. Am. Chem. Soc. 127, 1346–1347.

21. Fry, H. C., Lehmann, A., Sinks, L. E., Asselberghs, I., Tronin, A., Krishnan, V., Blasie, J. K., Clays, K., DeGrado, W. F., Saven, J.G. & Therien, M. J. (2013). Computational de novo design and characterization of a protein that selectively binds a highly hyperpolarizable abiological chromophore. J. Am. Chem. Soc. 135, 13914–13926.

22. Dantas, G., Corrent, C., Reichow, S. L., Havranek, J. J., Eletr, Z. M., Isern, N. G., Kuhlman, B., Varani, G., Merritt, E. A. & Baker, D. (2007). High-resolution structural and thermodynamic analysis of extreme stabilization of human procarboxypeptidase by computational protein design, J. Mol. Biol. 366, 1209–1221.

23. Ponder, J. W. & Richards, F. M. (1987). Tertiary templates for proteins: Use of packing criteria in the enumeration of allowed sequences for different structural classes. J. Mol. Biol. 193, 775–791.

24. Smith, C. A., Kortemme, T. (2008). Backrub-like backbone simulation recapitulates natural protein conformational variability and improves mutant side-chain prediction. J. Mol. Biol. 380, 742–756.

25. Lassila, J. K. (2010). Conformational diversity and computational enzyme design. Curr. Opin. Chem. Biol. 14, 676—682.

26. Gainza, P., Roberts, K. E. & Donald, B. R. (2012). Protein design using continuous rotamers. PLoS Comput. Biol. 8, e1002335.

27. Preiswerk, N., Beck, T., Schulz, J. D., Milovnk, P., Mayer, C., Siegel, J. B., Baker, D. & Hilvert, D. (2014). Impact of scaffold rigidity on the design and evolution of an artificial diels-alderase. Proc. Natl. Acad. Sci. USA 111, 8013–8018.

28. Pokala, N. & Handel, T. M. (2005). Energy Functions for Protein Design: Adjustment with Protein–Protein Complex Affinities, Models for the Unfolded State, and Negative Design of Solubility and Specificity. J. Mol. Biol. 347, 203—227.

29. Kellogg, E. H., Leaver-Fay, A., & Baker, D. (2011). Role of conformational sampling in computing mutation-induced changes in protein structure and stability. Proteins 79, 830–838.

30. Larson, S. M., England, J. L., Desjarlais, J. R. & Pande, V. S. (2002). Thoroughly sampling sequence space: Large-scale protein design of structural ensembles. Protein Sci. 11, 2804–2813.

31. Ollikainen, N. & Kortemme, T. (2013). Computational protein design quantifies structuralconstraints on amino acid covariation. PLoS Comput. Biol. 9, e1003313.

32. LaVallie, E. R., DiBlasio, E. A., Kovacic, S., Grant, K. L., Schendel, P. F. & McCoy, J. M. (1993). A thioredoxin gene fusion expression system that circumvents inclusion body formation in the E. coli cytoplasm. Biotechnology (N.Y.) 11, 187–193.

33. Benson, D. E., Wisz, M. S. & Hellinga, H. W. (2000). Rational design of nascent metalloenzymes. Proc. Natl. Acad. Sci. USA 97, 6292–6297.

34. Bolon, D. N. & Mayo, S. L. (2001) Enzyme-like proteins by computational design. Proc. Natl. Acad. Sci. USA 98, 14274–14279

35. Van de Streek, J. & Neumann, M. A. (2010). Validation of experimental molecular crystal structures with dispersion-corrected density functional theory calculations. Acta Crystallogr. Sect. B 66, 544–558.

36. Schneider, M., Fu, X. & Keating, A. E. (2009). X-ray vs. NMR structures as templates for computational protein design. Proteins 77, 97–110.

37. Kelley, R. F. & Richards, F. M. (1987) Replacement of proline-76 with alanine eliminates the slowest kinetic phase in thioredoxin folding. Biochemistry 26, 6765–6774.

38. Russel, M. & Model, P. (1986). The role of thioredoxin in filamentous phage assembly. Construction, isolation, and characterization of mutant thioredoxins. J. Biol. Chem. 261, 14997–5005.

39. Tanaka, J., Doi, N., Takashima, H. & Yanagawa, H. (2010). Comparative characterization of random-sequence proteins consisting of 5, 12, and 20 kinds of amino acids. Prot. Sci. 19, 786–795.

40. Neira, J. L., González, C., Toiron, C., de Prat-Gay, G. & Rico, M. (2001). Three-dimensional solution structure and stability of thioredoxin m from spinach. Biochemistry 40, 15246—15256.

41. Watters, A. L., Deka, P., Corrent, C., Callender, D., Varani, G., Sosnick, T., & Baker, D. (2007). The highly cooperative folding of small naturally occurring proteins is likely the result of natural selection. Cell 128, 613–624.

42. Piana, S., Lindorff-Larsen, K., & Shaw, D. E. (2013). Atomistic description of the folding of a dimeic protein. J. Phys. Chem. B 117, 12935–12942.

43. DeGrado, W. F., Raleigh, D. P., & Handel, T. (1991). De novo protein design: what are we learning? Curr. Opin. Struc. Biol. 1, 984 – 993.

44. Eklund, H., Gleason, F. K. & Holmgren, A. (1991). Structural and functional relations among thioredoxins of different species. Proteins 11, 13–28.

45. Langsetmo, K., Fuchs, J. A. & Woodward, C. (1991). The conserved, buried aspartic acid in oxidized Escherichia coli thioredoxin has a pK_a of 7.5. Its titration produces a related shift in global stability. Biochemistry 30, 7603—7609.

46. O'Meara, M. J., Leaver-Fay, A., Tyka, M. D., Stein, A., Houlihan, K., DiMaio, F., Bradley, P., Kortemme, T., Baker, D., Snoeyink, J. & Kuhlman, B. (2015). Combined covalent-electrostatic model of hydrogen bonding improves structure prediction with Rosetta. J. Chem. Theory Comput. 11, 609—622.

47. Nivón, L. G., Moretti, R. & Baker, D. (2013). A pareto-optimal refinement method for protein design scaffolds. PLoS ONE 8, e59004.

48. Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucl. Acids Res. 25, 3389—3402.

49. Cock, P. J. A., Antao, T., Chang, J. T., Chapman, B. A., Cox, C. J., Dalke, A., Friedberg, I., Hamelryck, T., Kauff, F., Wilczynski, B. & de Hoon, M. J. L. (2009). Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics 25, 1422—1423.

50. Joosten, R. P., Long, F., Murshudov, G. N. & Perrakis, A. (2014). The PDB_REDO server for macromolecular structure model optimization. IUCrJ 1, 213—220.

51. Berjanskii, M., Zhou, J., Liang, Y., Lin, G. & Wishart, D. S. (2012). Resolution-by-proxy: a simple measure for assessing and comparing the overall quality of NMR protein structures. J. Biomol. NMR 53, 167—180.

52. Dunbrack, R. L. & Cohen, F. E. (1997). Bayesian statistical analysis of protein side-chain rotamer preferences. Protein Sci. 6, 1661—1681.

53. Shapovalov, M. V. & Dunbrack, R. L. Jr. (2011). A smoothed backbone-dependent rotamer library for proteins derived from adaptive kernel density estimates and regressions. Structure 19, 844–858.

54. Larkin, M.A., Blackshields, G., Brown, N.P., Chenna, R., McGettigan, P.A., McWilliam, H., Valentin, F., Wallace, I.M., Wilm, A., Lopez, R., Thompson, J.D., Gibson, T.J. & Higgins, D.G. (2007). Clustal W and Clustal X version 2.0. Bioinformatics 23, 2947—2948.

55. Perrière, G. & Gouy, M. (1996). WWW-query: An on-line retrieval system for biological sequence banks. Biochimie 78, 364–369.

56. The PyMOL Molecular Graphics System, Version 1.4.1 Schrödinger, LLC.

57. R: A Language and Environment for Statistical Computing, R Core Team, Vienna, Austria, 2015, https://www.R-project.org.

58. Crooks, G. E., Hon, G., Chandonia, J. M. & Brenner, S. E. (2004). WebLogo: A sequence logo generator. Genome Res. 14, 1188—1190.

59. Olsson, M. H. M., Søndergaard, C. R., Rostkowski, M. & Jensen, J. H. (2011). Propka3: Consistent treatment of internal and surface residues in empirical pK_a predictions. J. Chem. Theory Comput. 7, 525–537.

60. Lauritsen, I., Willemoës, M., Jensen, K. F., Johansson, E. & Harris, P. (2011). Structure of the dimeric form of CTP synthase from Sulfolobus solfataricus. Acta Crystallogr. Sect. F 67, 201—208.

61. Hansen, M. R., Barr, E. W., Jensen, K. F., Willemoës, M., Grubmeyer, C. & Winther, J. R. (2014). Catalytic site interactions in yeast OMP synthase. Arch. Biochem. Biophys. 542 28–38.

62. Walters, J., Milam, S. L. & Clark, A. C. (2009). Chapter 1 practical approaches to protein folding and assembly: Spectroscopic strategies in thermodynamics and kinetics. Methods Enzymol. 455, 1—39.

63. Battye, T. G. G., Kontogiannis, L., Johnson, O., Powell, H.R. & Leslie A. G. W. (2011). IMosflm: a new graphical interface for diffraction-image processing with MOSFLM. Acta Crystallogr. Sect. D 67, 271—281.

64. Kabsch, W. (2010). XDS. Acta Crystallogr. Sect. D 66, 125—132.

65. Adams, P. D., Afonine, P. V., Bunkoczi, G., Chen, V.B., Davis, I. W., Echols, N., Headd, J. J., Hung, L. W., Kapral, G.J., Grosse-Kunstleve, R. W., McCoy, A. J., Moriarty, N. W., Oeffner, R., Read, R. J., Richardson, D. C., Richardson, J. S., Terwilliger, T. C. & Zwart, P. H. (2010). PHENIX: a comprehensive Python-based system for macromolecular structure solution. Acta Crystallogr. Sect. D 66, 213—221.

66. Emsley, P. & Cowtan, K. (2004). Coot: model-building tools for molecular graphics. Acta Crystallogr. Sect. D 60, 2126—2132.

Figure Legends

Fig. 1. Structural overlay of the eight experimentally determined template structures of thioredoxin showing the structural similarity.

Fig. 2. Wild-type sequences of the eight thioredoxin structures used as templates for computational design. The height of each letter is scaled according to frequency and the colors represent chemical properties.

Fig. 3. Sequence alignment of the 960 designed and 8 wild-type sequences. The length of the lines gives the fraction of pairwise sequence identity difference (scale bar at lower left). The 2D projection here focuses on close relationships. The nomenclature is given in the text. All the designs precisely cluster according to the template and one cluster represents most of the native thioredoxin sequences.

Fig. 4. Expression and solubility evaluation of 48 designed sequences. Each panel shows evaluations of one of the eight experimental templates. Protein levels in the pellet (grey bars) and supernatant (black bars) fractions were determined from anti-His_5 western blots by comparison to bands of control protein in known concentrations. 2 pmol corresponds to approximately 0.2 mg/L of E. coli culture at OD_{600} of 5. A missing bar indicates no detectable protein. Note the logarithmic scale, where <0.02 pmol and >200 pmol indicate the lower and upper detection limits, respectively.

Fig. 5. Structural and biophysical characterization of the two stable and monomeric designs, dF106 and dF414. (a) Size exclusion chromatography of dF106 kept on ice (red) or left at room temperature for three days (black), and freshly prepared dF414 (blue). (b) Far UV CD spectra of 11.1 µM dF106 in 25 mM sodium sulfate pH 4.8 and 7.6 µM dF414 in 25 mM sodium phosphate pH 7. The inset shows the high-tension voltage, indicating reasonable accuracy over the entire spectrum. (c) Chemical denaturation of dF106 and dF414, with the data normalized using a fit to a three-state model (raw data in Fig. S4). The fits are shown merely as a guide for the eye

as the simple model does not explain the data. (d) Representative spectra for the unfolding of dF106 in GuHCl showing the large shift in λ_{max}.

Fig. 6. Comparing the X-ray structure of dF106 to the template and the design model. (a) Composite simulated annealing omit map (2mFo-DFc) contoured at 2 σ. The high consistency between the model and the omit map demonstrates little to no model bias in the final structure. (b) Overlay of the template 1FB0 (green), design model (red), and design X-ray structure (blue). Pro 37 is show shown with sticks. (c) Examples of reproduction of core side chain conformations.

Fig. 7. Sequences of the experimentally characterized designs compared to the wild-type 1FB0. A sequence logo constructed form the 18 designs that result in any soluble albeit unstable protein (black bars in Fig. 4, excluding dF106 and dF104) is shown below the template 1FBO, and the two stable and monomeric designs, dF106 and dF414. A sequence logo constructed from the 12 sequences that expressed but were insoluble (gray bar and no black bar in Fig. 4) is shown at the bottom. The letters (amino acids) are colored according to chemical properties and heights of the soluble and insoluble sequences are scaled according to frequency.

Fig. 8. Reproduction of 1FB0 wild-type identities in computational designs based on 1FB0 and the correlation with evolutionary conservation. For each position, the reproduction frequency of the wild-type identity in designs based on the experimental template (black bars) and 5 geometry-optimized templates (white bars) is shown. The wild-type amino acid (letter below bars) is colored if naturally conserved with respect to structure (red) or function (green).[44]

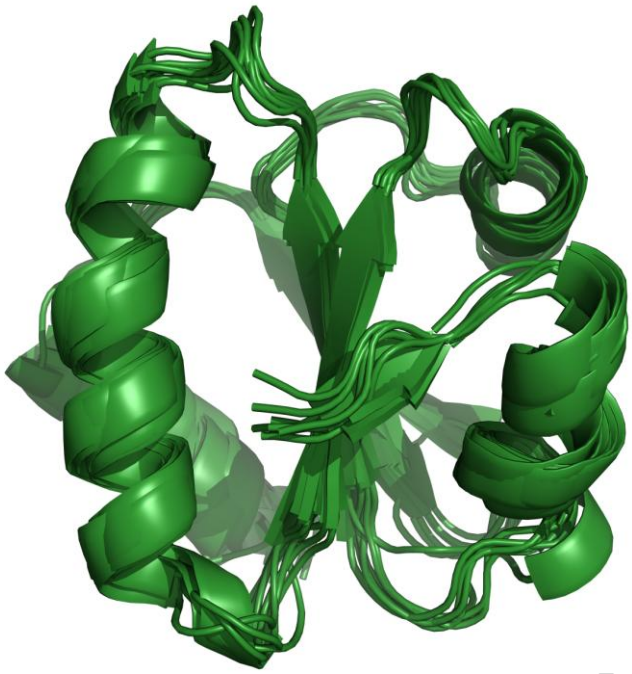


Figure 1

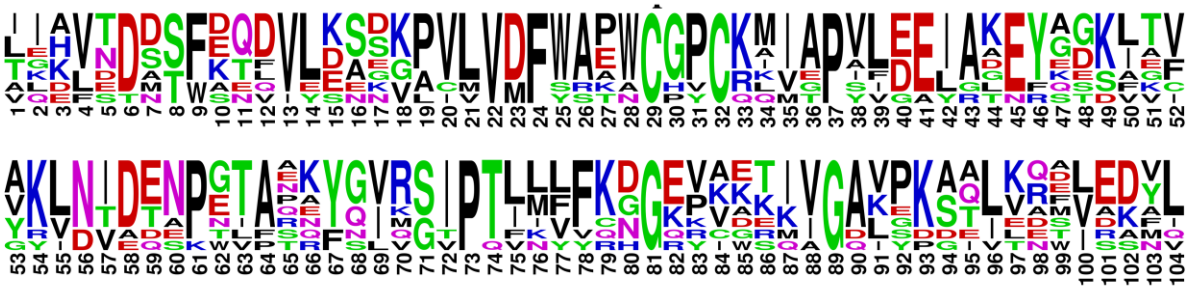


Figure 2

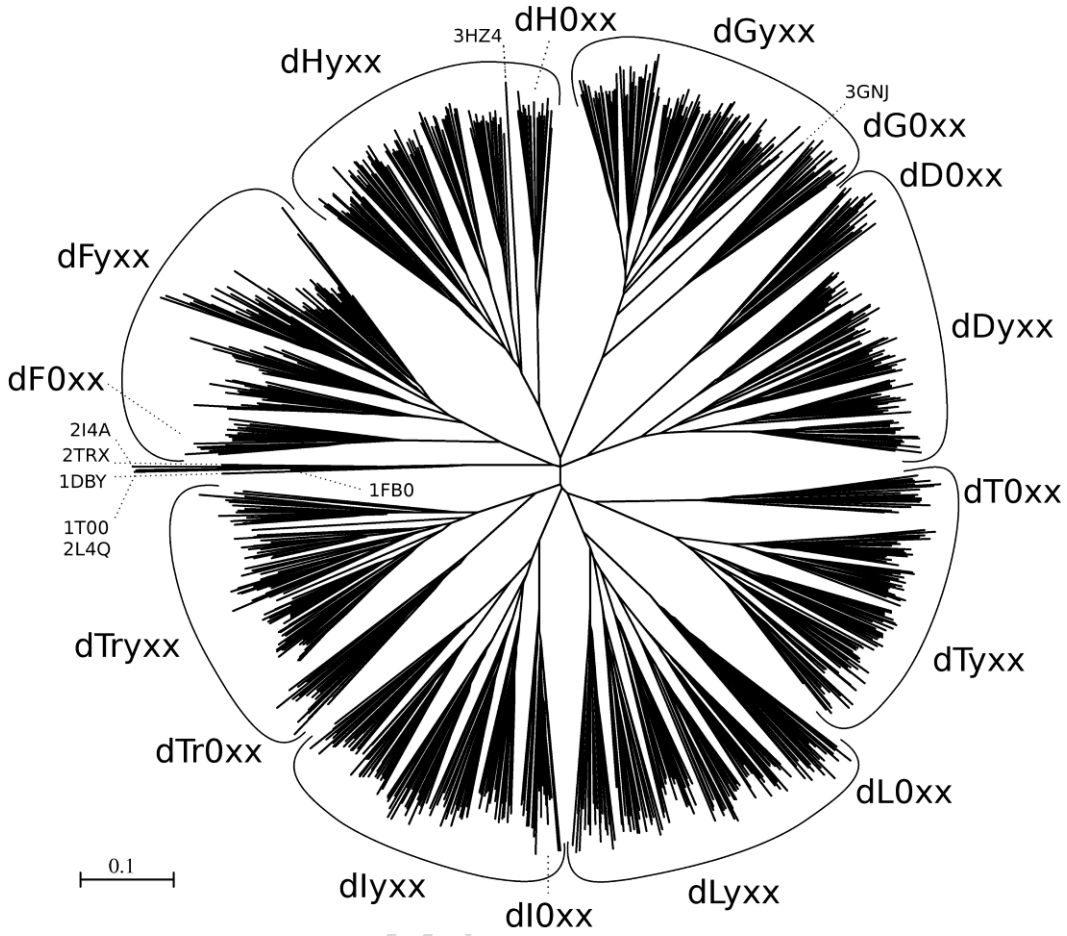


Figure 3

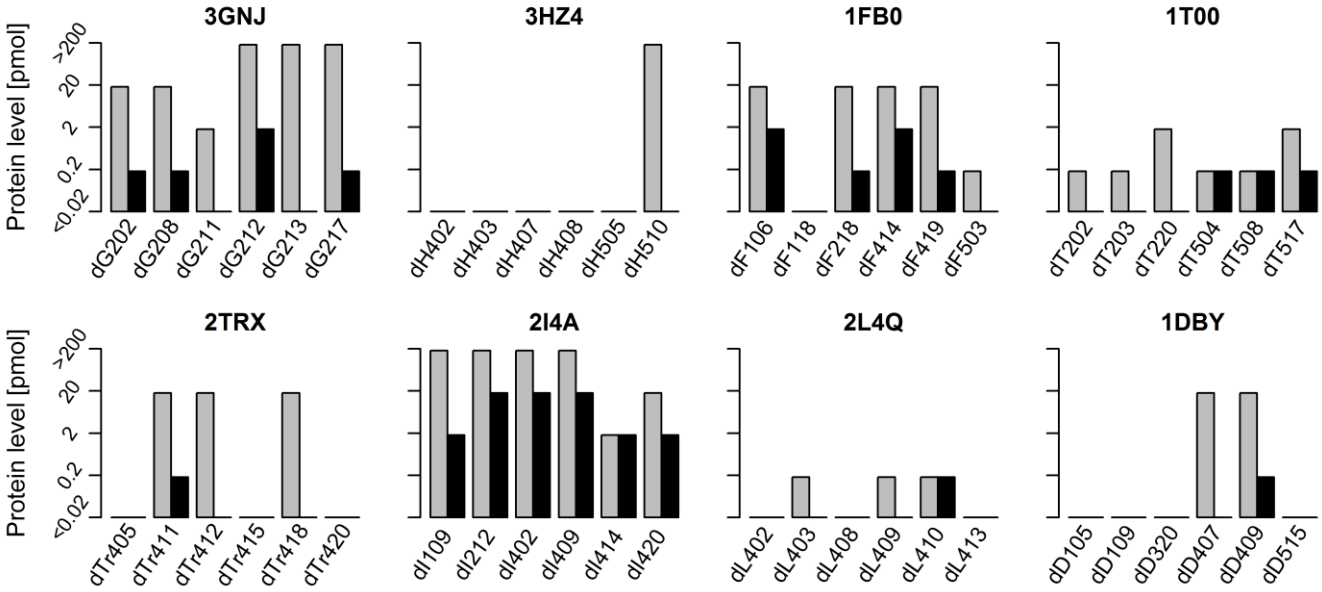


Figure 4

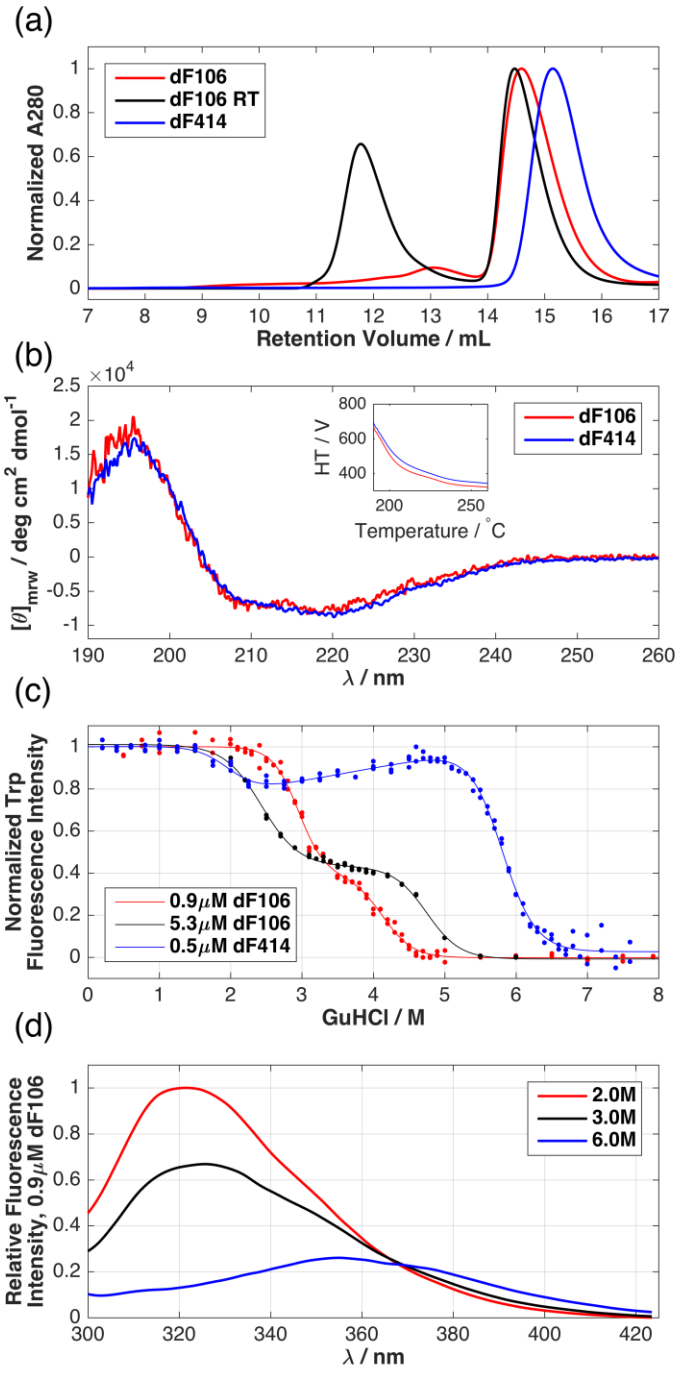


Figure 5

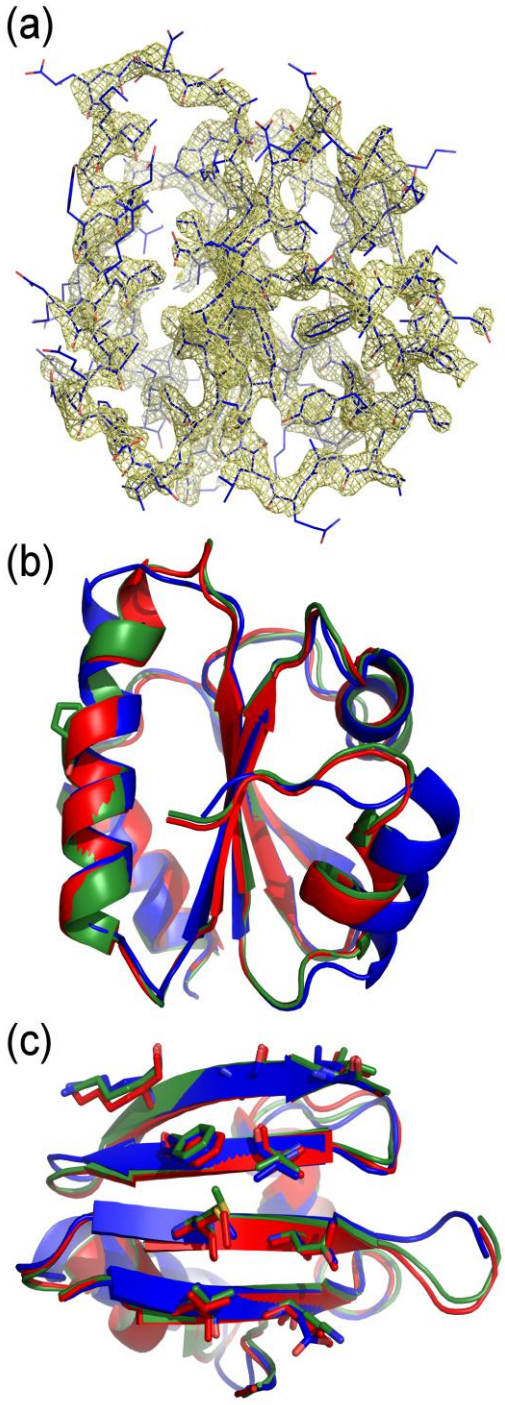


Figure 6

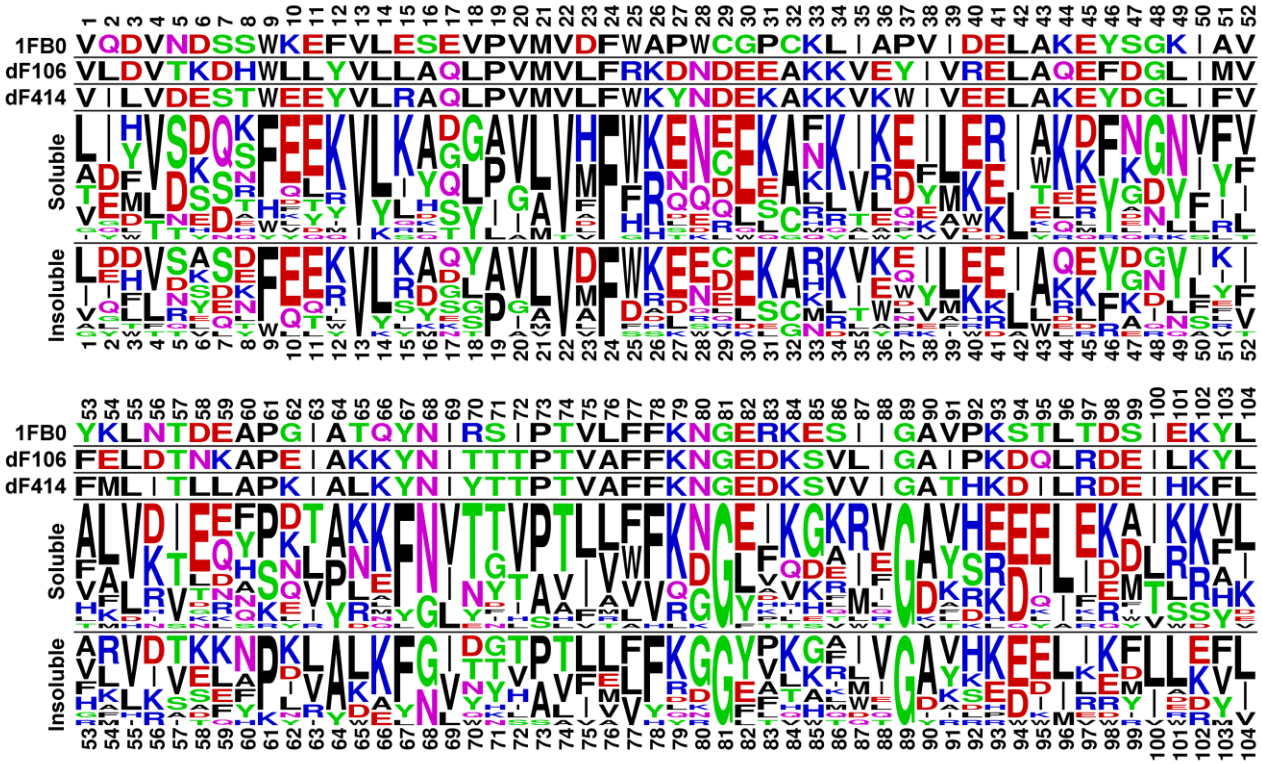


Figure 7

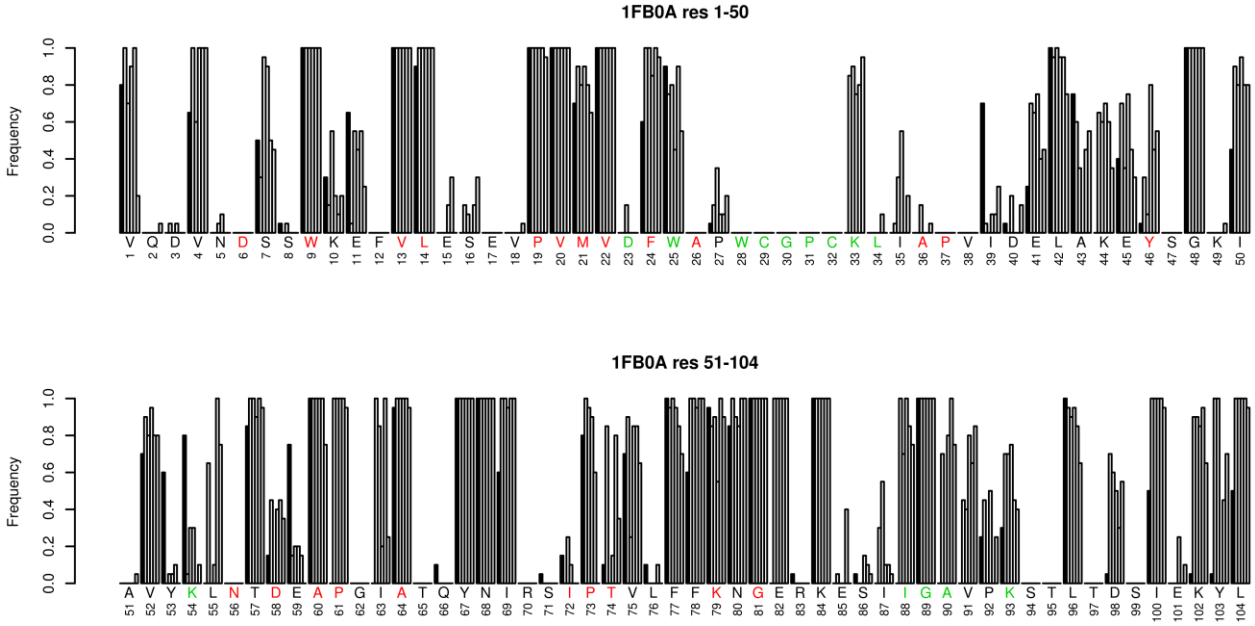


Figure 8

Table 1. Thioredoxin structures used as design templates

PDB ID	2I4A	1T00	2TRX	3GNJ	1FB0	3HZ4	2L4Q	1DBY
Residues	4–107	5–108	4–107	3–106	9–112	6–109	9–112	3–106
Resolution (Å)	1.00	1.51	1.68	1.99	2.26	2.30	NMR	NMR
Organism	*Acetobacter aceti*	*Streptomyces coelicolor A3(2)*	*E. coli*	*Desulfitobacterium hafniense DCB-2*	*Spinacia oleracea*	*Methanosarcina mazei*	*Mycobacterium tuberculosis*	*Chlamydomonas reinhardtii*

Table 2. Template designability evaluation by geometry optimization and core side-chain repacking

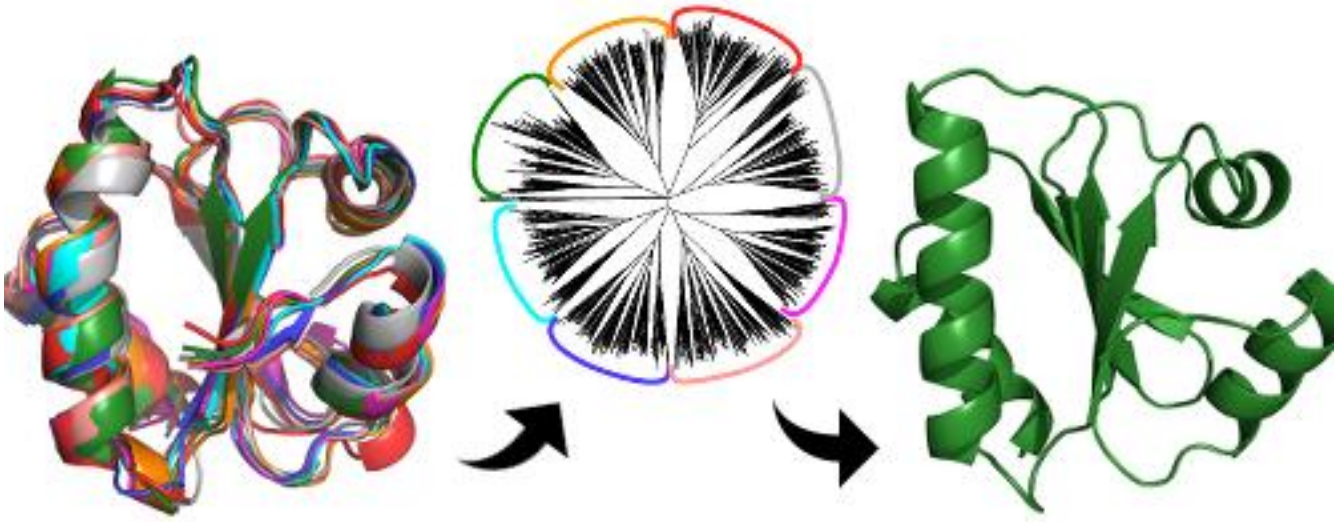
Template	2I4A	1T00	2TRX	3GNJ	1FB0	3HZ4	2L4Q	1DBY
Resolution (Å)	1.00	1.51	1.68	1.99	2.26	2.30	NMR	NMR
Average energy (REU)	−226	−230	−228	−243	−236	−230	−217	−198
Average distortion (Å)	0.7	0.7	0.5	0.6	0.7	0.9	0.9	1.0
χ_1 reproduction (%)	88	90	98	98	100	98	76	95
χ_1 reproduction[a] (%)	100	100	100	100	100	100	99	100

[a]Geometry-optimized templates

Table 3. Average energy of position 23 in Rosetta energy units (REU)

	Asp			Leu		
1T00	N	Packer energy	Relax energy	N	Packer energy	Relax energy
A	85	-1.6	-2.0	0	-	-
B	0	-	-	100	-1.4	-2.4
C	95	-1.7	-2.0	0	-	-
D	15	-	-2.4	0	-	-
E	0	-	-	97	-2.4	-2.8
2TRX	Asp			Leu		
A	92	-2.1	-2.3	2	-2.3	-2.9
B	0	-	-	100	-2.2	-2.6
C	89	-2.1	-2.3	8	-2.0	-2.6
D	0	-	-	9	-	-2.9
E	0	-	-	100	-2.4	-2.7

A: Original settings. B: Original settings with position 23 fixed as Leu. C: Maximal rotamer library at position 23. D: Talaris2013 energy function (relax using score12 energy function for comparison). E: D23L template with original settings.



Graphical abstract

Highlights

- Computational protein design methods suffer from low success rates.
- An automated redesign of the thioredoxin fold was validated by an X-ray structure.
- Computational design is found to be highly sensitive to the backbone template.
- Thorough geometry optimization prior to design can result in artifacts.
- Using more templates can improve the overall chance of success.